MORPH-II: A Proposed Subsetting Scheme

Participants: K. Kempfert, J. Fabish, K. Park, and R. Towner
Mentors: Y. Wang, C. Chen, and T. Kling
NSF-REU Site at UNC Wilmington, Summer 2017

Abstract

In this paper, we propose a new subsetting scheme for the longitudinal face aging database MORPH-II. Our subsetting scheme is intended to overcome the unbalanced racial and gender distributions of MORPH-II, while ensuring independence between training and testing sets. Our subsetting scheme can be used for various face analysis tasks, including gender classification, age prediction, and race classification.

1 Introduction

MORPH is one of the largest publicly available longitudinal face databases [5]. Since its first release in 2006, it has been cited by over 500 publications, as determined by our Google Scholar search. Multiple versions of MORPH have been released, but for our subsetting scheme we use the 2008 MORPH-II non-commercial release (which will be referred to as MORPH2008 here). MORPH2008 includes over 55,000 longitudinal mugshots, taken between 2003 and late 2007. For each image, the following metadata is included: subject ID number, picture number, date of birth, date of arrest, race, gender, age, time since last arrest, and image filename. Because of its size, its longitudinal span, and its inclusion of relevant metadata, MORPH2008 has been used for a variety of race, gender, and age face imaging tasks. In addition to MORPH2008, we also briefly address a previous version of MORPH-II, which has been used by face imaging researchers in the past. Such discussion can be found in the following section.

2 Background

Our subsetting scheme is motivated by Guo's and Mu's work in [2, 3, 4], which is done on a previous version (with unknown date) of the MORPH-II noncommercial dataset, which here we will refer to as MORPHpre. MORPHpre and MORPH2008 are very similar, but there are some minor differences. Most images are the same, but there are 2 images in MORPH2008 that are not included in MORPHpre. MORPHpre includes an additional race category **I**ndian, in addition to the 5 races in MORPH2008: **W**hite, **B**lack, **A**sian, **H**ispanic, **O**ther. Beyond the specific differences mentioned, MORPH2008 is an updated, cleaner version of MORPHpre. This information is briefly summarized in Table 1.

For comparative purposes, the experimental design by Guo and Mu from [2, 3, 4] will be summarized first. About 66.75% of images in MORPHpre are of black males, while only about 4.72% of images are of white females. Guo and Mu utilize a subsetting scheme to overcome such disproportionate distributions of racial and gender groups in MORPHpre. They denote the whole dataset as W and randomly select a subset S of 21,060 white and black faces of both genders from ages 16 to 67. Only white and black subjects are included in S, because other races have too few images for use in the training set. In S, half the images are white, and the other half are black. There are

three times as many males as females in S. R denotes the set of remaining images, $W \setminus S$. In R, there are approximately 34,000 images including both genders, all 6 races, and ages 16-77 years. These subsets are summarized in Figure 1.

Table 1: Differences between MORPHpre and MORPH2008

	Number of Images	**Race Categories**	**Other Qualities**
MORPHpre [2, 3, 4]	55,132	W,B,H,A,I,O	older version
MORPH2008	55,134	W,B,H,A,O	newer version; cleaner

S is further divided equally into $S1$ and $S2$, with 10,530 images for $S1$ and $S2$, respectively. $S1$ and $S2$ are used alternately for training and testing purposes. First, $S1$ is used for training and $W \setminus S1$ for testing. Then $S2$ is used for training and $W \setminus S2$ for testing. This process is similar to 2-fold cross-validation, but $S1$ and $S2$ are not obtained by a random partition of the images in S; instead, $S1$ and $S2$ are controlled to be as similar as possible.

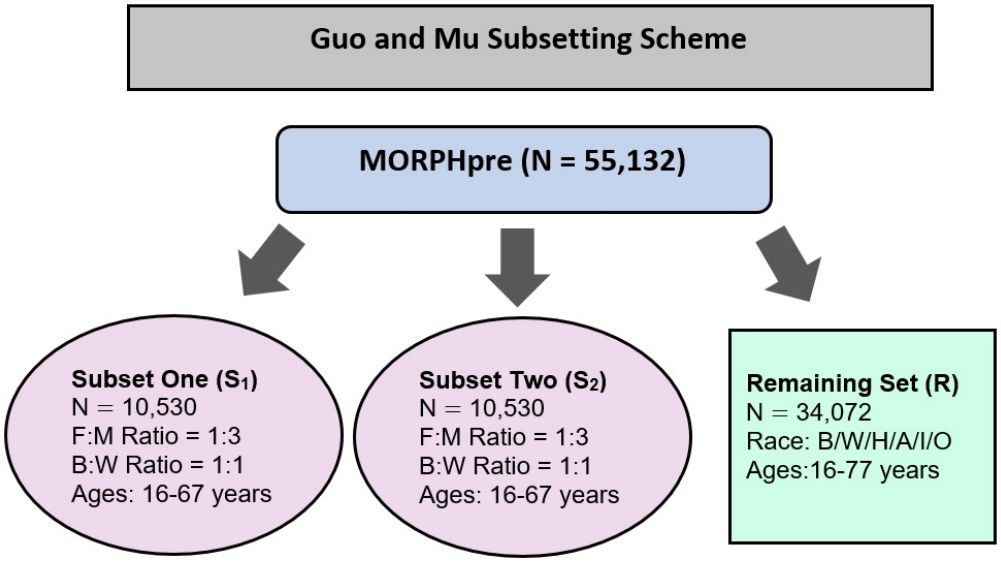

Figure 1: This flowchart represents Guo's and Mu's subsetting scheme in [2, 3, 4].

In $S1$ and $S2$, Guo and Mu ensure equivalent age distributions within gender. For example, the set of white females in S1, $S1_{WF}$, has an identical age distribution to the set of black females in S1, $S1_{BF}$. The equivalence of age distributions is summarized in Figure 2. Through this subsetting scheme, Guo and Mu form training sets that are more proportionate in gender and race. Additionally, they guarantee identically distributed ages within each gender class.

However, the training and testing sets are not necessarily independent. It has been found that a number of subjects belong to both $S1$ and $S2$ and, in some cases, even to R. For this reason, there may be information leakage between these sets. Consequently, any models built and validated according to this subsetting scheme may suffer from inflated accuracy rates (or deflated error rates).

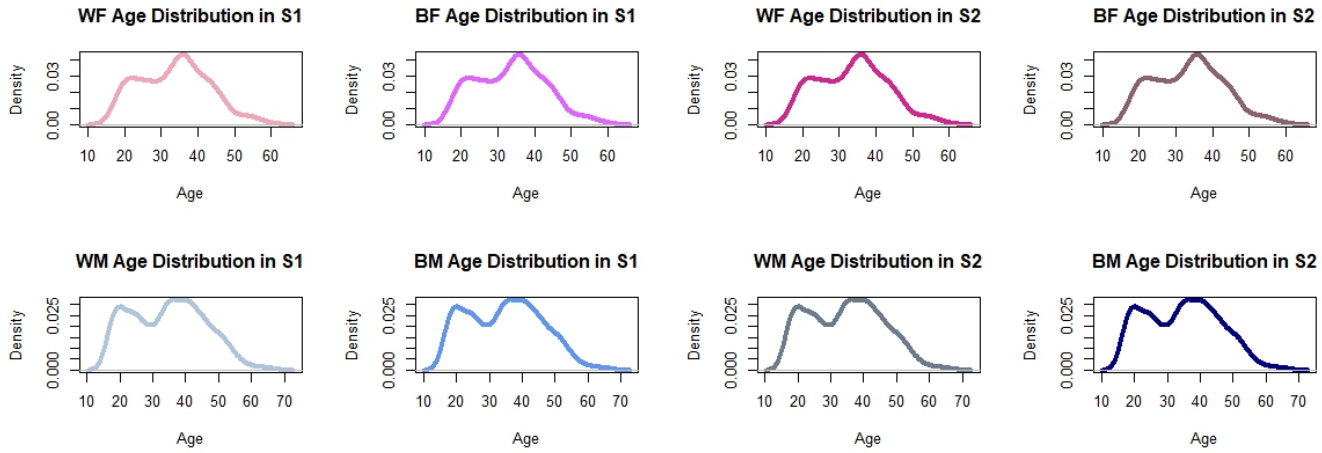

Figure 2: Age Distributions for subsets $S1$ and $S2$ in S are shown for Guo's and Mu's subsetting scheme [2, 3, 4]. On the top row of the figure, it is shown that $S1_{WF}$, $S1_{BF}$, $S2_{WF}$, and $S2_{BF}$ have identical age distributions. The bottom row of the figure depicts the identical age distributions for $S1_{WM}$, $S1_{BM}$, $S2_{WM}$, and $S2_{BM}$.

3 New Subsetting Scheme

3.1 Development

Like Guo and Mu, we only consider the white and black races for the training sets, since the number of images for other racial groups is too small. In our subsetting scheme, we seek to retain Guo's and Mu's ratios of white:black and male:female images, while ensuring independence between training and testing sets. We also prioritize randomization, aiming to create many candidate subsets from which to choose. These possible subsets can be used for comparative purposes in the future; models could be built and validated on different subsets, and the results could be averaged or compared. Additionally, we use MORPH2008, since it is the most up-to-date version of MORPH-II. Further, we generate our subsets from a version of MORPH2008 that has been more thoroughly cleaned.

The motivation for cleaning MORPH2008 is the detection of various inconsistencies in gender, race, and birthdates among repeat offenders in MORPH2008. Such inconsistencies are validated and corrected. Note that a new variable *age dec* is also created to represent the exact age in decimal of the subject pictured in each image. Hereafter, those age values in decimal will be used in subsetting and future facial demographical study, since they are less biased than integer-valued ages. The decimal age values are also advantageous due to their improved continuity, which is essential as an assumption for the nonparametric tests discussed in a later section. More details on the cleaning of the MORPH2008 dataset can be found in [1]. The cleaned version of MORPH2008 from [1] is used for the generation of subsets in this study.

For ease of comparison, we attempt to be as consistent as possible with the set notation in [2, 3, 4]. Let W be the Whole cleaned dataset (*morphII_go_for_age*), S be the main training/validation set, and R be the remaining set. We divide S into $S1$ and $S2$, such that $S1$ and $S2$ have the same number of images. We fix the ratios of white:black images to be 1:1 and male:female images to be 3:1.

Because white females are the smallest race-gender combination, we include all 2,570 white females in S. We randomly allocate each white female subject to either $S1$ or $S2$ exclusively,

according to the constraint that the total number of white female images in S1 is equal to the number of white female images in S2:

$$|S1_{WF}| = 1285 = |S2_{WF}|.$$

Note that all white females are included in S, hence none are included in R.

For the other race-gender categories (black females, white males, and black males), we include only a portion of their images in S, while the remainder goes in R. For black females, we randomly allocate a subset of subjects to $S1$ and an exclusive subset of subjects to $S2$, such that the total number of black female images in $S1$ is equal to the total number of black female images in S2, as well as equal to the total number of white female images in S_i, $i = 1, 2$. The images pertaining to any remaining black female subjects are sent to R.

$$|S1_{BF}| = 1285 = |S2_{BF}| = |Si_{WF}|, i = 1, 2.$$

For white males, we randomly allocate some subjects to $S1$ and other distinct white male subjects to $S2$, such that 3 times the number of white female images are in S1. The number of white male images in $S1$ is also set to be equal to the number of white male images in $S2$. Any remaining white males' images are sent to R.

$$|S1_{WM}| = 3855 = |S2_{WM}| = 3|Si_{WF}| = 3|Si_{BF}|, i = 1, 2.$$

The same process is repeated for black males, so that there are equal numbers of black male images within S1 and S2. The number of black male images is equal to the number of white male images, and other equalities hold too:

$$|S1_{BM}| = 3855 = |S2_{BM}| = |Si_{WM}| = 3|Si_{WF}| = 3|Si_{BF}|, i = 1, 2.$$

In this way, we ensure independence between $S1$, $S2$, and R. There is no expected information leakage between the training and testing sets. However, it should be clarified that observations within each set are not independent. For each subject s_j in some set Ω, all of s_j's images are in Ω. Hence, some observations within each set Ω are correlated with each other. Though we are able to guarantee independence between training and testing sets, at this time we have no satisfactory solution to the issue of correlated images within each set. This is an issue inherent to longitudinal data.

3.2 Implementation

We implement our subsetting scheme in the statistical software R. We iterate through various values of k ($k = 1, 2,...$). Subsets are randomly generated for each value of k, with a random seed of k set anytime randomization is invoked. In this way, numerous candidate subsets are created.

Among the candidate subsets, we seek those with similar age distributions. We obtain the age distributions of images in $S1$ and $S2$. Then for each value of k, we perform both the Anderson-Darling (AD) and Kolmogorov-Smirnov (KS) tests on those distributions. The hypotheses for both tests are as follows:

$$H_o = S1_{age} \text{ has the same distribution as } S2_{age}$$
$$H_a = S1_{age} \text{ does not have the same distribution as } S2_{age}$$

We use the P-Values of both tests to identify the best subsets. High P-Values indicate $S1$ and $S2$ have similar age distributions for a particular seed k. Hence, we use the P-Values for these

nonparametric tests as metrics for judging suitable subsets. In this context, the P-Values are not to be interpreted as clear probabilities, for the following reasons: not all assumptions for the tests are met (since some observations within each set are dependent) and the significance level cannot be defined appropriately when an indefinite number of tests are made. We believe our unconventional use of P-Values is valid here, since we are not attempting to make any probability statements based off them.

Using these criteria, we identify random seed number $k = 42$ as one which produces satisfactory subsets. The P-Values for the KS and AD tests are 0.657 and 0.652, respectively. The statistical summaries below further indicate the suitability of these subsets. We do not guarantee that random seed 42 produces the global optimum results, but it is found to be satisfactory for our purposes.

In Tables 2, 3, and 4, we include basic information pertaining to the subsets generated by the random seed 42. We intend the subsets W, S, and R to be used as Guo and Mu did in [2, 3, 4]: first the model should be trained on $S1$ and tested on $W \setminus S1$, then the model should be trained on $S2$ and tested on $W \setminus S2$. Then two sets of results can be summarized.

Table 2 shows the number of images in subsets by race and gender, while Table 3 gives the number of distinct subjects in subsets by race and gender. More detailed information on the different races is summarized in Table 4, with the number of images and the number of distinct subjects for the additional race groups in the remaining subset R. In Tables 2 and 3, **d** denotes different race subjects (**H**ispanic, **A**sian, or **O**ther).

Table 2: Number of Images in Subsets by Race and Gender

	WF	**BF**	**WM**	**BM**	**dF**	**dM**	**Overall**	**F**	**M**
S1	1,285	1,285	3855	3,855	0	0	10,280	2570	7,710
S2	1,285	1,285	3,855	3,855	0	0	10,280	2,570	7,710
R	0	3,150	220	28,980	144	1,850	34,344	3,294	31,050
Overall	2,570	5,720	7,930	36,690	144	1,850	54,904	8,434	46,470

Table 3: Number of Distinct Subjects in Subsets by Race and Gender

	WF	**BF**	**WM**	**BM**	**dF**	**dM**	**Overall**	**F**	**M**
S1	311	332	1,005	948	0	0	2,596	643	1,953
S2	313	336	988	943	0	0	2,580	649	1,931
R	0	809	55	6,899	40	568	8,371	849	7,522
Overall	624	1,477	2,048	8,790	40	568	13,547	2,141	11,406

Table 4: Additional Race Groups in Remaining Subset R

	HF	**AF**	**OF**	**HM**	**AM**	**OM**	**Overall**	**F**	**M**
Subjects in R	28	4	8	502	47	19	608	40	568
Images in R	99	13	32	1,646	140	64	1,994	144	1,850

Additional graphical and numerical summaries are presented in Figures 3, 4, 5, and Table 5.

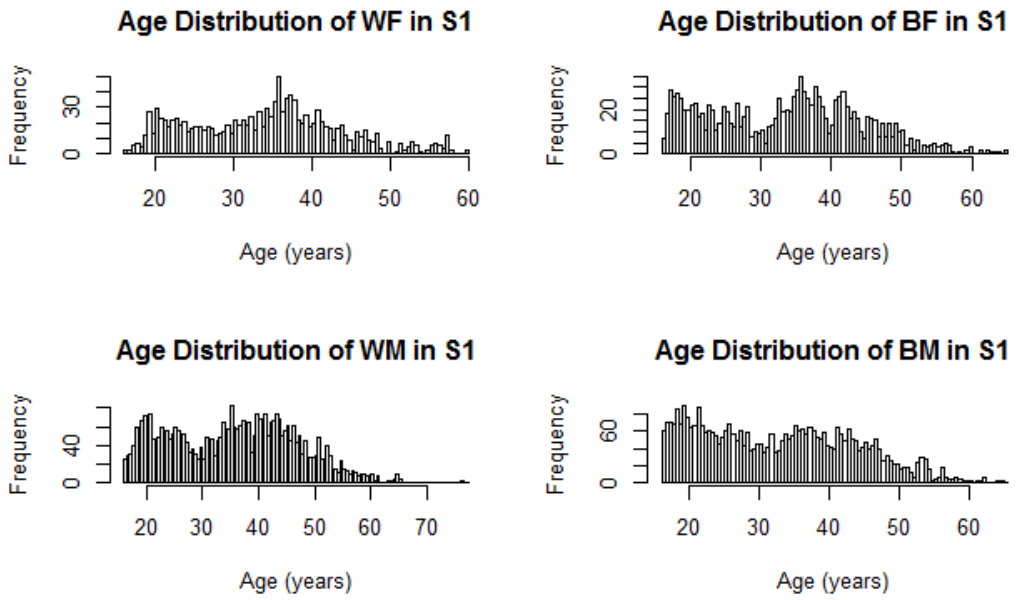

Figure 3: For random seed 42, the observed age histograms in $S1$ are displayed.

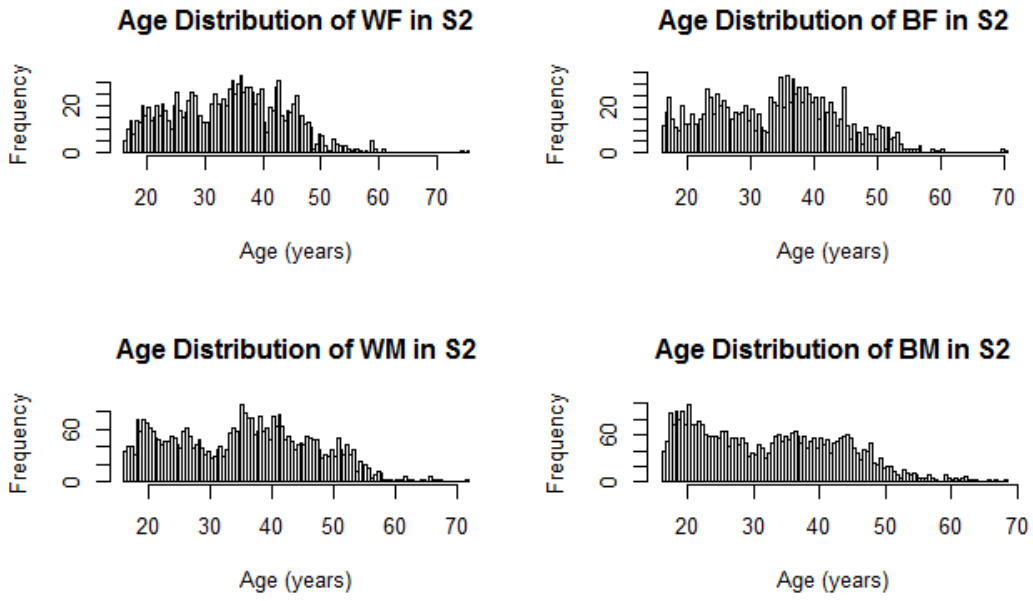

Figure 4: For random seed 42, the observed age histograms in $S2$ are displayed.

Figure 3 displays the observed age histograms in $S1$. It is shown that all the gender-race combinations in $S1$ have similar, right-skewed age distributions. Any differences in distribution here seem minor and unlikely to significantly affect gender or race classification in future experiments. Figure 4 presents the observed age histograms in $S2$. Based on the plots, all the gender-race combinations in $S2$ seem to be similarly distributed. Further, we see that the age distributions in $S1$ are not much different than the age distributions in $S2$. We do not expect any deviations in age distribution between sets to negatively impact classification in a significant way.

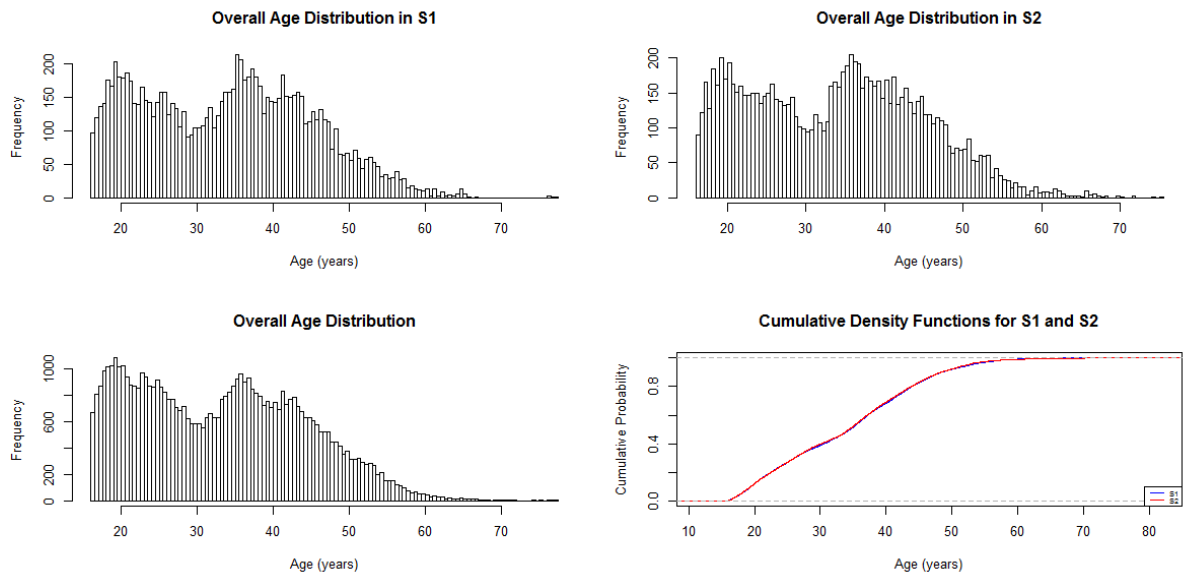

Figure 5: The age distributions for images in $S1$, $S2$, and the W dataset are depicted.

The age distributions for images in $S1$, $S2$, and the W dataset are depicted in Figure 5. It is shown that all three histograms are right-skewed with a roughly bimodal structure, indicating that $S1$ and $S2$ have been chosen successfully; the age distributions of images in the subsets $S1$ and $S2$ are close to the overall age distribution of images in W. The final plot shows the ECDFs of $S1$ and $S2$. It is difficult to distinguish the densities corresponding to each subset, since departures are so minor. This aligns with our expectations, because the P-Values for the KS and AD tests were quite large (approximately 0.65 for each).

In Table 5, the 5-Number Summary, as well as mean and standard deviation, are given for age in $S1$, $S2$, and W. The statistical summaries are nearly identical, further confirming these subsets' balanced age distributions.

Table 5: Numerical Summary of Age in Sets

	Min.	**Q1**	**Median**	**Q3**	**Max**	**Mean**	**SD**
S1	16.003	24.296	34.495	42.185	77.196	34.041	10.957
S2	16.005	24.370	34.371	42.014	75.421	33.926	10.908
W	16	23.369	33.091	41.422	77.196	33.019	10.950

All numerical and graphical summaries we consider here indicate the suitability of the subsets generated from random seed 42. These subsets are expected to yield good results for a variety of face imaging tasks, including gender and race classification, as well as age regression.

4 Conclusion

In our paper, we propose a subsetting scheme of the MORPH-II aging database. Our scheme is inspired by the work of Guo and Mu, but we do make some changes. Most notably, we maintain the racial and gender proportions of Guo and Mu, while ensuring independence between training and testing sets. Our approach is also novel in its generation of various candidate subsets, which are selected based off nonparametric goodness of fit tests KS and AD. We present one suitable choice

of subsets for a random seed of 42, but the generation of other subsets from the random seeds k are recommended for comparative purposes in the future. For any models built and tested using the subsetting scheme proposed in this study, it is expected the estimates of test error or accuracy can be less biased. Our subsetting scheme can be used for face imaging tasks involving gender, race, and age.

5 Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Numbers DMS-1659288. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] G. Bingham, B. Yip, M. Ferguson, C. Nansalo, C. Chen, Y. Wang, and T. Kling. *MORPH-II: Inconsistencies and Cleaning Whitepaper*, 2017. http://libres.uncg.edu/ir/uncw/f/wangy2017-1.pdf.

[2] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 71–78. IEEE, 2010.

[3] G. Guo and G. Mu. A study of large-scale ethnicity estimation with gender and age variations. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 79–86. IEEE, 2010.

[4] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pages 657–664. IEEE, 2011.

[5] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.